Code Summarization: Results Achieved with Zero-shot When Prompting Detailed Instructions

LLM	No detailed instructions		With instructions	
	#	%	#	%
CodeLlama 7B	1	0.17%	1	0.18%
CodeLlama 13B	23	3.87%	0	0.00%
CodeLlama 34B	2	0.34%	43	7.57%
GPT-3.5-turbo	0	0.00%	6	1.06%
GPT-4-turbo	0	0.00%	0	0.00%

Table 1: Code Summarization: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment. Comparison between the scenarios where the models are prompted with (right) and without (left) detailed instructions about how to judge each summary quality aspect.

Quality aspect	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
Content adequacy	-0.11	0.05	-0.11	0.05	0.52
Conciseness	-0.26	-0.82	0.02	0.13	0.20
Fluency	-0.05	-0.04	0.15	0.25	0.23

Table 2: Code Summarization: Krippendorff agreement between the LLMs judgments and the median human rating.

	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
Content Adequacy									
human	2.89	3.59	3.73	4.11	4.67	2.97	-	-	-
CL 7B	1.34	0.61	0.40	0.26	-0.42	1.24	*** (L)	*** (L)	(N)
CL 13B	1.04	0.36	0.23	-0.17	-0.67	0.91	* (S)	*** (M)	** (S)
CL 34B	1.62	1.11	0.80	0.77	0.30	1.70	(N)	(N)	*** (M)
GPT 3.5	1.13	0.55	0.43	-0.08	-0.55	1.02	*** (S)	*** (M)	*** (L)
GPT 4	0.35	-0.23	-0.16	-0.30	-0.56	0.15	*** (S)	*** (M)	*** (L)
Conciseness									
human	4.53	4.74	4.57	4.92	4.82	4.80	-	-	-
CL 7B	-0.38	-0.70	-0.55	-0.70	-0.74	-0.71	** (S)	** (S)	** (S)
CL 13B	-1.72	-1.86	-1.74	-2.10	-1.91	-2.04	(N)	(N)	(S)
CL 34B	-0.52	-0.42	-0.63	-0.35	-0.36	-0.88	(N)	* (S)	(N)
GPT 3.5	0.25	0.01	0.14	0.07	0.03	-0.17	(N)	(N)	** (S)
GPT 4	-0.25	-0.33	-0.20	-0.04	0.11	-0.24	*** (S)	*** (S)	*** (S)
Fluency&Understandability									
human	4.40	4.68	4.80	4.67	4.83	3.71	-	-	-
CL 7B	0.18	-0.12	-0.45	0.07	-0.18	1.06	** (S)	(N)	*** (M)
CL 13B	-0.52	-0.74	-0.88	-0.80	-0.88	0.10	(N)	(N)	*** (M)
CL 34B	0.36	0.13	-0.09	0.33	0.17	1.12	*** (S)	*** (S)	*** (L)
GPT 3.5	0.28	0.08	-0.08	0.08	0.01	0.84	(N)	(N)	*** (M)
GPT 4	0.28	0.08	-0.02	0.28	0.12	0.88	(N)	(N)	*** (M)

Adjusted *p*-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 3: Code Summarization: Average of differences between the LLM judgments and the ground truth (*i.e.,* the median of the score given by the three human raters). Last three columns report adj. *p*-value and effect size when comparing the judgments each LLM gave to summaries it generated against those it gave when judging summaries (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

Human	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
GPT-4 (4.67)	GPT-3.5 (4.34)	GPT-4 (4.00)	GPT-4 (4.62)	CL 34B (4.15)	GPT-4 (4.10)
GPT-3.5 (4.11)	GPT-4 (4.25)	CL 34B (3.96)	GPT-3.5 (4.54)	CL 13B (4.14)	GPT-3.5 (3.81)
CL 34B (3.73)	CL 7B (4.23)	CL 13B (3.95)	CL 7B (4.52)	GPT-4 (4.03)	CL 34B (3.56)
CL 13B (3.59)	human (4.21)	GPT-3.5 (3.94)	CL 34B (4.41)	CL 7B (4.02)	CL 13B (3.36)
human (2.97)	CL 13B (4.20)	CL 7B (3.94)	CL 13B (4.28)	GPT-3.5 (3.99)	CL 7B (3.24)
CL 7B (2.89)	CL 34B (4.12)	human (3.89)	human (3.80)	human (3.87)	human (3.12)

Table 4: Content adequacy: ranking of the generators of summaries according to each judge, including both humans and LLMs.

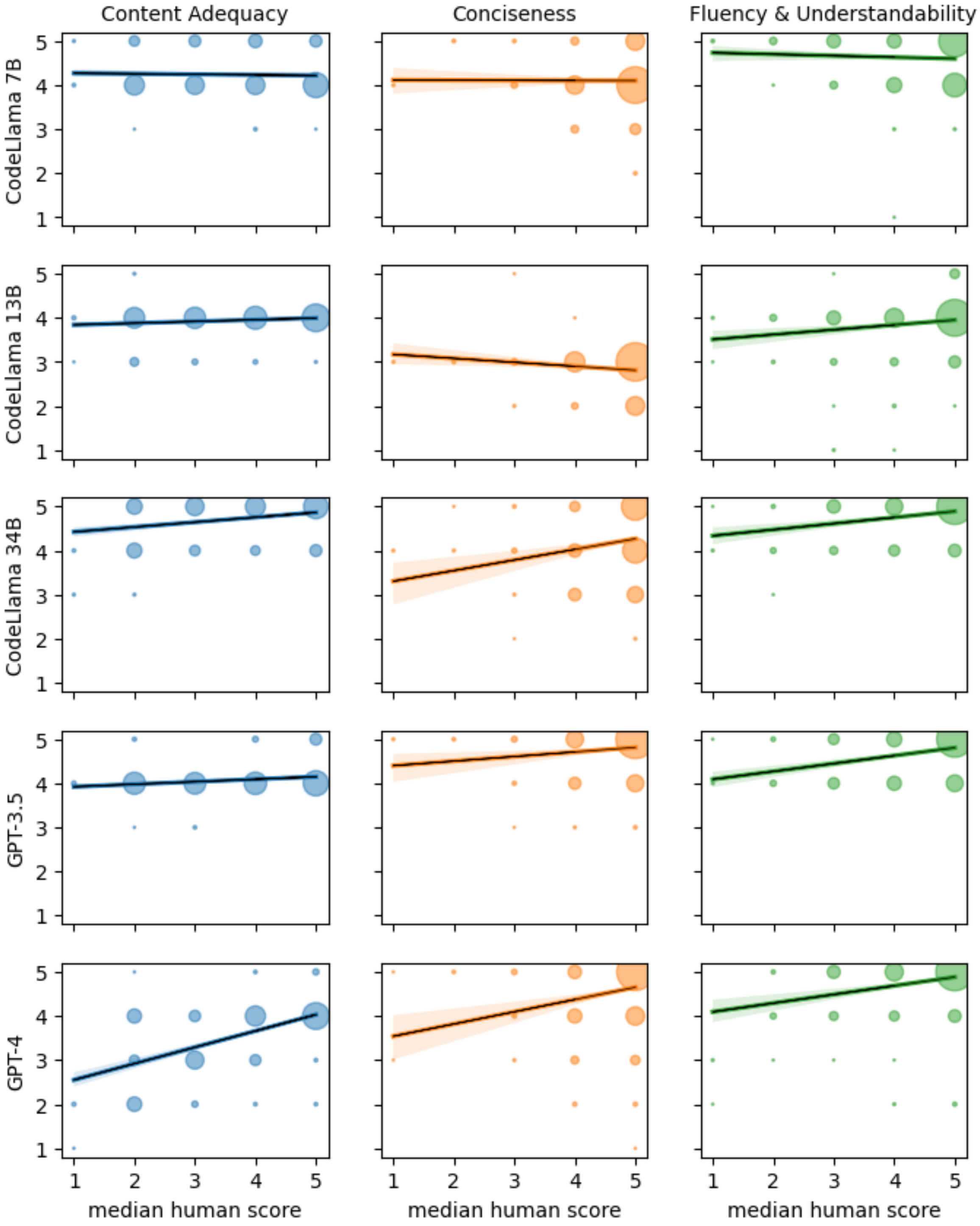

Figure 1: Code Summarization: Scatterplots relating human to LLM judgments.